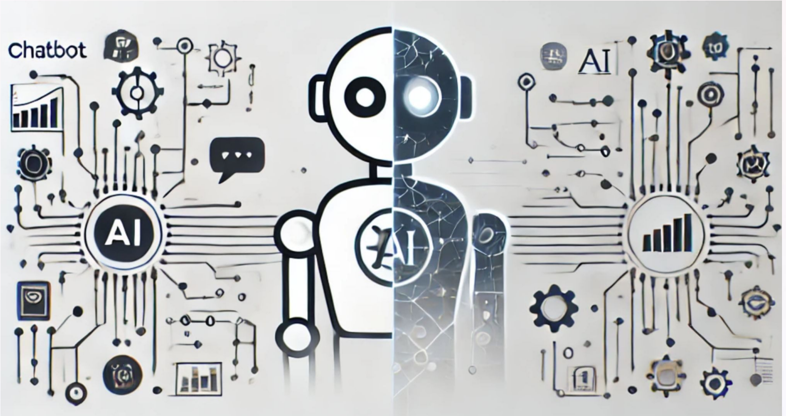
AI-Powered Rails: From Chatbots to Intelligent Assistants

Szymon Kurcab
KRUG #3 2024

Meet Szymon

- Started working with Ruby and Rails in 2004
- First KRUG meetups in 2005/2006
- Co-owner of kina.krakow.pl/repertuary.pl
- Interested in AI since watching Terminator I
  … ok, maybe closer to 2020/2021
- Following closely AI space since Nov 2022
- Currently working as Head of AI Labs @ Tropic
- Contributing to ruby-openai gem
- Recently reviewed Obie Fernandez new book: "Patterns of Application Development Using AI"

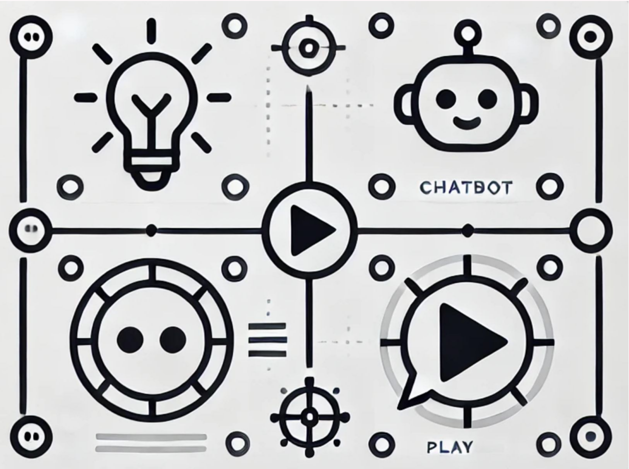


Links:

- @simonx1 on X/Twitter
- szymonk on LinkedIn
- Ruby AI Builders Discord server

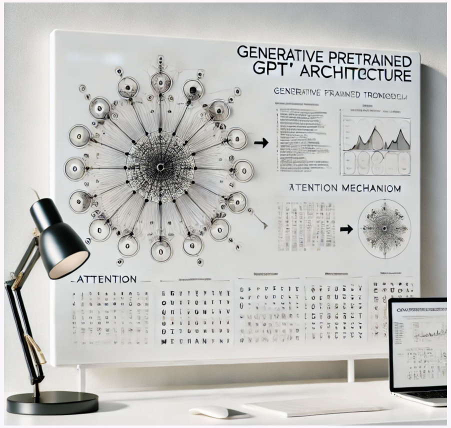
Agenda

- Basic GenAI Concepts
- Chatbot vs AI Assistant vs AI Agent
- Implementing AI Assistant in Rails
- Live demos



GenAI Concepts

Generative Pretrained Transformer (GPT) architecture



Generative Pretrained Transformer (GPT) architecture is a type of artificial intelligence model designed to understand and generate human-like text. Here's what each part means:

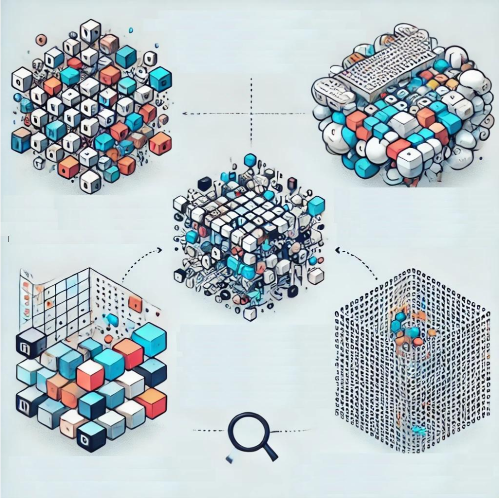
- **Generative**: It can produce new content, like writing sentences or paragraphs that read as if a human wrote them.
- **Pretrained**: The model has already been trained on a large amount of text data (like books and articles), so it has learned grammar, facts about the world, and some reasoning abilities.
- **Transformer**: This is the underlying technology that allows the model to process and generate text efficiently. Transformers use a mechanism called "attention"* to understand the context and relationships between words in a sentence.**

* Attention Is All You Need 20217

** o1-preview definition

GenAI Concepts

- Tokens
- Text embedding
- Latent Space
- Context Window
- Retrieval Augmented Generation (RAG)



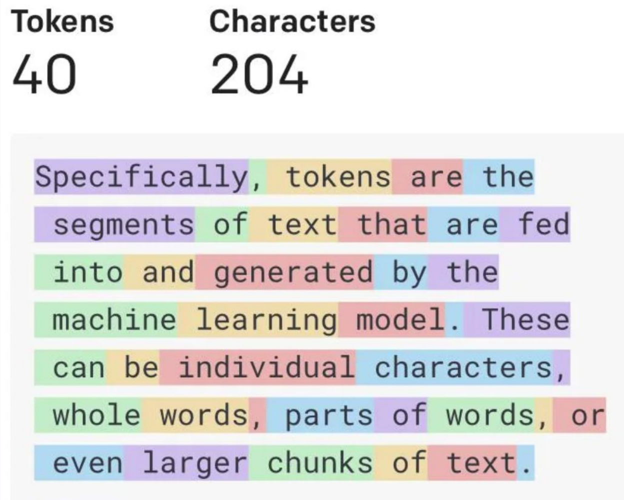
GenAI Concepts

Tokens: The Building Blocks of AI Language

- Tokens are like the alphabet of AI language understanding
- They can be parts of words, whole words, or even punctuation
- Example: "I love Ruby!" might be tokenized as ["I", "love", "Ru", "by", "!"]

AI models build understanding from these basic token units.

Why it matters: The number of tokens affects how much an AI can process at once and how much it costs to use.



Tokens
40

Characters
204

Specifically, tokens are the segments of text that are fed into and generated by the machine learning model. These can be individual characters, whole words, parts of words, or even larger chunks of text.

GenAI Concepts

Text Embeddings: The AI's Secret Language

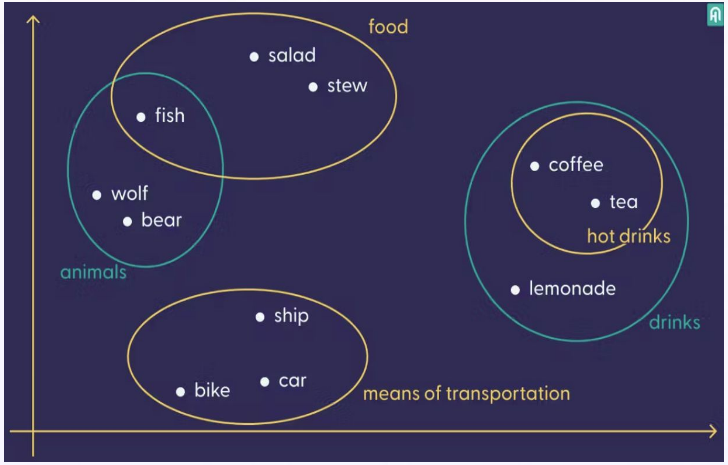
- Text embeddings are numerical representations of words or phrases
- They capture semantic meaning in a way computers can process
- Similar words or concepts have similar embeddings

Imagine a magical translator that doesn't just translate words between human languages, but instead translates them into a universal "meaning language" made of numbers. This is what embeddings do for AI.

Example:

- The word "ruby" might have an embedding like [0.2, -0.5, 0.7, ...]
- In this number space, "ruby" would be closer to "programming" than to "jewelry"

Why it matters: Embeddings allow AI to understand relationships between words and concepts, enabling more nuanced language understanding and generation.



*deepset.ai

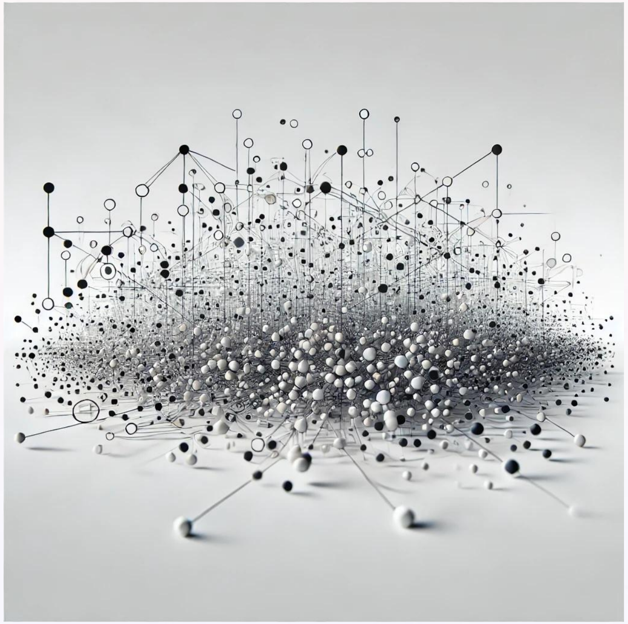
GenAI Concepts

Latent Space: The AI's Imagination

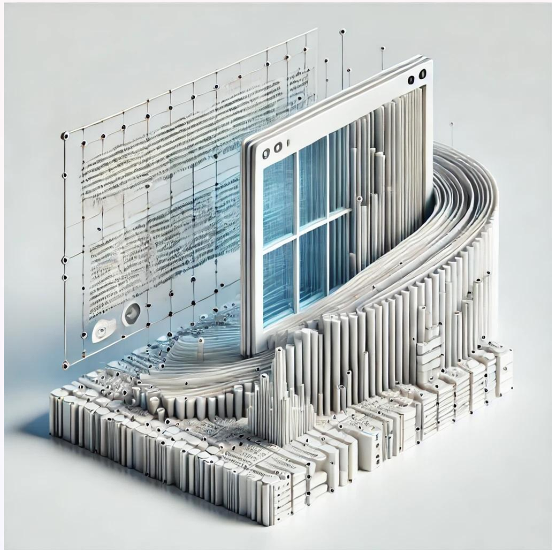
- Latent space is a complex, multi-dimensional space where AI represents concepts
- It's where the AI "understands" and connects ideas
- Similar concepts are closer together in this space

Imagine a vast library where every book represents a concept. The AI organizes this library so that similar books are near each other. The way it navigates this library to find and connect ideas is its "latent space."

Why it matters: This is how AI can understand context and generate relevant responses.



GenAI Concepts



Context Windows: How Much Can AI "See"?

- The context window is how much text the AI can consider at once
- Larger windows allow for understanding of longer documents or conversations
- But larger windows also mean more processing time and cost

Why it matters: The size of the context window affects the AI's ability to maintain coherence over long conversations or analyze large documents.

GenAI Concepts

RAG: Giving AI Access to External Knowledge

- RAG combines the AI's built-in knowledge with external information
- It allows AI to use up-to-date or specialized information
- Enhances accuracy and relevance of AI responses

Think of RAG as giving the AI the ability to quickly "Google" facts while it's thinking. It's not limited to just what it learned during training.

Why it matters: RAG helps create AI systems that are more accurate and can handle specialized or current topics better.

Chatbot vs AI Assistant vs AI Agent

Chatbot

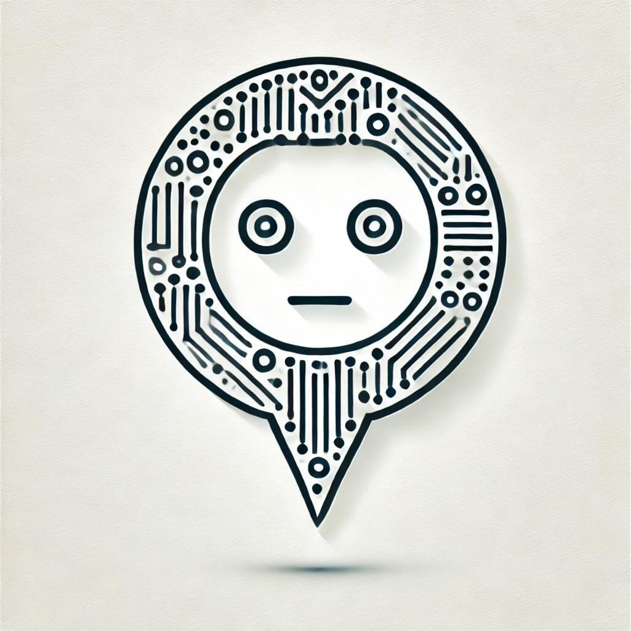
A program designed to simulate conversation with human users, especially over the Internet.

Key Characteristics:

- Often rule-based or using simple pattern matching
- Typically handles simple, predefined tasks
- Limited context awareness
- Usually stateless (doesn't remember previous interactions)

Analogy: Like a very basic customer service rep with a script.

Example in Rails: A simple bot that answers FAQs on a website.



Chatbot vs AI Assistant vs AI Agent

AI Assistant

A more advanced conversational AI system that can understand context and perform a variety of tasks.

Key Characteristics:

- Uses natural language processing (NLP) for better understanding
- Can handle more complex, open-ended queries
- Maintains context within a conversation
- Often has some form of memory or user profile awareness
- Can integrate with various services to perform actions

Analogy: Like a knowledgeable personal assistant who can handle a wide range of tasks.

Example in Rails: Tropic AI Request Assistant - to be demoed

Chatbot vs AI Assistant vs AI Agent

AI Agent

An autonomous AI system that can perceive its environment, make decisions, and take actions to achieve specific goals.

Key Characteristics:

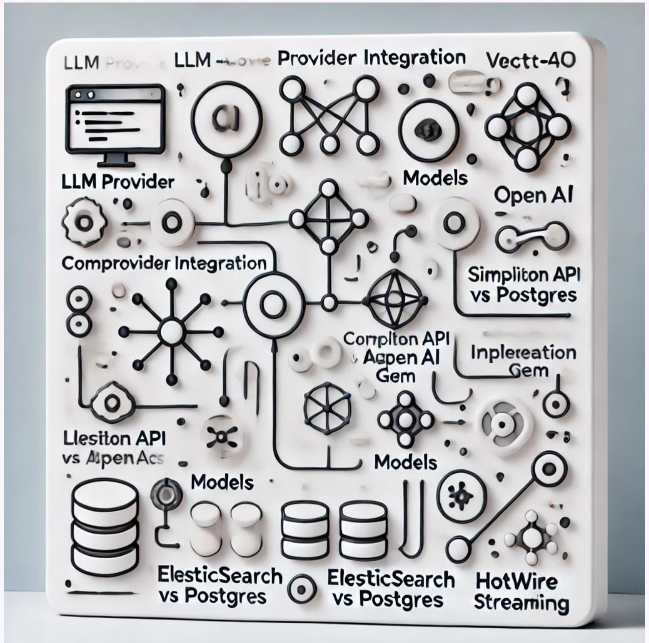
- Has defined goals and can make decisions to achieve them
- Can operate autonomously without constant human input
- Often uses advanced AI techniques like reinforcement learning
- Can interact with its environment and learn from outcomes
- May collaborate with other agents or humans

Example in Rails: An CLI tool that automatically fixes broken specs. To be demoed.

Implementing a Chatbot in Rails

Key ingredients

- LLM provider integration
  - gpt-4o/gpt-4o-mini model + ruby-openai gem
  - Completion API vs Assistant API
- Vector DB storage
  - ElasticSearch vs Postgres vector storage
- RAG implementation
  - Simplistic vs Advanced RAG (for e.g. Contextual Retrieval)
- UI/UX
  - Hotwire Turbo Streaming (WS) and the AI model's output



Demo

QnA

Thank you